One Way Analysis of Variance

INTD8065 Data Analysis for Cancer Research
María-Eglée Pérez and Luis Raúl Pericchi

Contents of this class

1. One Way Analysis of Variance Model
   - Model
   - Hypothesis Testing
   - Example
   - Fitting One Way ANOVA in *R*
   - Comparing Individual Means
   - Effect of departures from the hypothesis

One Way Analysis of Variance Model

Consider k groups of observations, and let n_i be the number of observations in group i.

y_{11}	y_{21}	$\cdots$	y_{k1}
y_{12}	y_{22}	$\cdots$	y_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{1n_1}	y_{2n_2}	$\cdots$	y_{kn_k}

We want to know if the means of the groups are equal.

Model

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

Errors are supposed to be independent and normally distributed with mean 0 and common variance σ^2

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I).$$

This is clearly a linear model. Solving the normal equations, the following least squares estimators for the μ_i's are obtained:

$$\hat{\mu}_i = \bar{y}_{i.} = \frac{\displaystyle\sum_{j=1}^{n_i} y_{ij}}{n_i}$$

Hypothesis Testing

For determining if there are differences between the groups, the following hypothesis can be tested

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \text{ vs } H_1 : \text{some } \mu_i \text{ different}$$

This hypothesis test is equivalent to comparing the models

$$\begin{aligned} y_{ij} &= \mu + \varepsilon_{ij} \\ y_{ij} &= \mu_i + \varepsilon_{ij} \end{aligned}$$

These models are nested, and they can be compared using the ANOVA table corresponding to the test of significance of the model.

Source	df	SS	MS	F
Treatments	$k-1$	$SSTr = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$MSTr = \frac{SSTr}{k-1}$	$\frac{MSTr}{MSE}$
Error	$n-k$	$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$		

$(n = \sum n_i)$

We reject H_0 when $F > F^{\alpha}_{k-1,n-k}$.

Example (Cuckoos Eggs in Nests of Other Species)

That cuckoo eggs were peculiar to the locality where found was already known in 1892. A study by E.B. Chance in 1940 called *The Truth About the Cuckoo* demonstrated that cuckoos return year after year to the same territory and lay their eggs in the nests of a particular host species. Further, cuckoos appear to mate only within their territory. Therefore, geographical sub-species are developed, each with a dominant foster-parent species, and natural selection has ensured the survival of cuckoos most fitted to lay eggs that would be adopted by a particular foster-parent species.

The data is drawn from the work of O.M. Latter in 1902. We want to decide if this data support differences in the lengths of the cuckoos according to foster-parent species.

```
> cuckoos.frm <- read.csv("cuckoosred.csv")
> names(cuckoos.frm)
[1] "egg.length"   "host.species"
> attach(cuckoos.frm)
> class(egg.length)
[1] "numeric"
> class(host.species)
[1] "factor"
> plot(host.species,egg.length)
```

Fitting One Way ANOVA in *R*

An analysis of variance model can be fitted in *R* using command aov

```
> cuckoos.mod1 = aov(egg.length˜host.species)
> summary(cuckoos.mod1)
             Df Sum Sq Mean Sq F value   Pr(>F)
host.species  4  41.07  10.268   12.62 8.74e-08 ***
Residuals    70  56.96   0.814
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

According to this ANOVA table, the data gives enough evidence to reject the null hypothesis of equality of the means. This means that the lengths of cuckoos eggs vary according the the surrogate parents species.

We can still use the command `lm` for this family of models, and results obtained are equivalent

```
> anova(lm(egg.length˜host.species))
Analysis of Variance Table

Response: egg.length
             Df Sum Sq Mean Sq F value    Pr(>F)
host.species  4 41.071 10.2677  12.619 8.744e-08 ***
Residuals    70 56.956  0.8137
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The following reparametrization is frequently used

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad , \quad \begin{array}{l} i = 1, \ldots, k \\ j = 1, \ldots, n_i \end{array}$$

Even though this model is equivalent to the first model we established ($\mu_i = \mu + \alpha_i$), it has more parameters than groups, and the normal equations will give have infinite solutions (the parameters of the model are *unidentifiable*)

To overcome this problem, it is necessary to impose a restriction on the α_i's. Some usual restrictions are

- $\sum_{i=1}^{k} n_i \alpha_i = 0$: the α_i's represent deviations of the mean of each group from the general mean.
- $\alpha_1 = 0$: $\alpha_i,\ i = 2, \ldots, k$ the α_i's represent deviations of the mean of each group with respect to the first group.

Default restriction in R is $\alpha_1 = 0$; it can be changed using `options`.

```
> levels(host.species)
[1] "Hedge.Sparrow" "Pied.Wagtail"  "Robin"
[2] "Tree.Pipit"    "Wren"
> coef(cuckoos.mod1)
             (Intercept) host.speciesPied.Wagtail
             23.12142857              -0.21809524
       host.speciesRobin   host.speciesTree.Pipit
             -0.54642857              -0.03142857
        host.speciesWren
             -1.99142857
> coef(cuckoos.mod1)[2:5]+23.12142857
host.speciesPied.Wagtail        host.speciesRobin
                22.90333                 22.57500
  host.speciesTree.Pipit         host.speciesWren
                23.09000                 21.13000
```

Changing the restriction

```
> options(contrasts=c("contr.sum","contr.poly"))
> cuckoos.mod2 = aov(egg.length~host.species)
> coef(cuckoos.mod2)
    (Intercept) host.species1 host.species2 host.species3
    22.56395238    0.55747619    0.33938095    0.01104762
  host.species4
     0.52604762
> coef(cuckoos.mod2)[2:5]+22.56395238
host.species1 host.species2 host.species3 host.species4
     23.12143      22.90333      22.57500      23.09000
> 22.56395238-sum(coef(cuckoos.mod2)[2:5])
[1] 21.13
```

In both cases, we obtain the same means for the different species of surrogate parents:

Hedge Sparrow	23.12	Robin	22.58
Tree Pipit	23.09	Pied Wagtail	22.90
Wren	21.13		

Comparing Individual Means

If the hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ is rejected, we will be interested in testing the hypotheses

$$H_0 : \mu_i = \mu_j \text{ vs } H_1 : \mu_i \neq \mu_j$$

There are several methods for performing these comparisons

- **Least Significant Difference test (LSD test)**
  If H_0 is true, it can be shown that

$$t = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim t_{n-k}$$

So we will reject H_0 with a fixed level α when H_0 si $|t| > t_{n-k}^{\alpha/2}$

This is equivalent to rejecting H_0 when

$$|\bar{y}_i - \bar{y}_j| > t_{n-k}^{\alpha/2}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

In *R*, the means of the groups can be compared using Least Significant Difference using the command `pairwise.t.test`

```
> pairwise.t.test(egg.length,host.species,p.adjust.method="none")

Pairwise comparisons using t tests with pooled SD

data:  egg.length and host.species

             Hedge.Sparrow Pied.Wagtail Robin   Tree.Pipit
Pied.Wagtail 0.52          -            -       -
Robin        0.10          0.31         -       -
Tree.Pipit   0.93          0.57         0.12    -
Wren         1.0e-07       9.2e-07      3.1e-05 9.6e-08

P value adjustment method: none
```

Everytime we perform a big number of group comparisons, the probability of obtaining no rejections (even when all means are equal) decreases very fast. This means that, when the number of groups is "large", the LSD method will reject true equality hypotheses just by chance, and globally, the type I error level α is larger than the nominal value (for a funny illustration of this phenomenon, see `http://xkcd.com/882/`)
To overcome this problem, several strategies can be adopted.

- **Adjusting p-values for controlling Type I error**
  - **Bonferroni method**

    The Type I error level α us adjusted for each individual test by

    $$\alpha = \frac{\alpha_T}{C}$$

    where α_T is the desired global level and C is the total number of comparisons. It is equivalent to multiply each p-value by C. This is a very conservative method, specially when the number of comparisons is very large. (Decreasing $\alpha \Rightarrow$ increasing β).

```
> pairwise.t.test(egg.length,host.species,p.adjust.method="bonferroni")

Pairwise comparisons using t tests with pooled SD

data:  egg.length and host.species

             Hedge.Sparrow Pied.Wagtail Robin   Tree.Pipit
Pied.Wagtail 1.00000       -            -       -
Robin        1.00000       1.00000      -       -
Tree.Pipit   1.00000       1.00000      1.00000 -
Wren         1.0e-06       9.2e-06      0.00031 9.6e-07

P value adjustment method: bonferroni
```

- **Sequential Bonferroni (Holm 1979)**
  The C test statistics (or the p-values) are ranked from largest to smallest and the smallest p-value is tested at α/c, the next at $\alpha/(c-1)$, the next at $\alpha/(c-2)$, etc.
  This procedure provides more power for individual tests and is recommended for any situation in which the Bonferroni adjustment is applicable.

```
> pairwise.t.test(egg.length,host.species,p.adjust.method="holm")

Pairwise comparisons using t tests with pooled SD

data:  egg.length and host.species

             Hedge.Sparrow Pied.Wagtail Robin   Tree.Pipit
Pied.Wagtail 1.00000       -            -       -
Robin        0.61405       1.00000      -       -
Tree.Pipit   1.00000       1.00000      0.61405 -
Wren         9.6e-07       7.4e-06      0.00022 9.6e-07

P value adjustment method: holm
```

- **Hochberg's procedure**
  The procedure is similar to Holm's, but works in reverse. The largest p-value is tested at α, rejecting all other tests if this one is significant. If not significant, the next largest is tested against $\alpha/2$, and so on. Hochberg's procedure is slightly more powerful that Holm's

```
> pairwise.t.test(egg.length,host.species,p.adjust.method="hochberg")

Pairwise comparisons using t tests with pooled SD

data:  egg.length and host.species

             Hedge.Sparrow Pied.Wagtail Robin   Tree.Pipit
Pied.Wagtail 0.92557       -            -       -
Robin        0.58330       0.92557      -       -
Tree.Pipit   0.92557       0.92557      0.58330 -
Wren         9.0e-07       7.4e-06      0.00022 9.0e-07

P value adjustment method: hochberg
```

- **Tukey's HSD test**
  (HSD=Honestly Significant Difference)
  John Tukey introduced intervals based on the range of the sample means rather than the individual differences. The intervals returned by this function are based on his Studentized Range Statistic. Technically the intervals constructed in this way would only apply to balanced designs where there are the same number of observations made at each level of the factor. The function in *R* incorporates an adjustment for sample size that produces sensible intervals for mildly unbalanced designs.

```
> TukeyHSD(cuckoos.mod1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = egg.length ~ host.species)

$host.species
                                  diff        lwr        upr     p adj
Pied.Wagtail-Hedge.Sparrow -0.21809524 -1.1567188  0.7205284 0.9659722
Robin-Hedge.Sparrow        -0.54642857 -1.4707834  0.3779263 0.4679638
Tree.Pipit-Hedge.Sparrow   -0.03142857 -0.9700522  0.9071950 0.9999821
Wren-Hedge.Sparrow         -1.99142857 -2.9300522 -1.0528050 0.0000010
Robin-Pied.Wagtail         -0.32833333 -1.2361065  0.5794398 0.8486583
Tree.Pipit-Pied.Wagtail     0.18666667 -0.7356318  1.1089651 0.9794348
Wren-Pied.Wagtail          -1.77333333 -2.6956318 -0.8510349 0.0000090
Tree.Pipit-Robin            0.51500000 -0.3927732  1.4227732 0.5096254
Wren-Robin                 -1.44500000 -2.3527732 -0.5372268 0.0002910
Wren-Tree.Pipit            -1.96000000 -2.8822985 -1.0377015 0.0000009
```

This intervals can be plotted for making a graphical comparison

```
> plot(TukeyHSD(cuckoos.mod1))
```

When the F test allows to reject the null hypothesis $H_0 : \alpha_i = 0, \quad i = 1, \ldots, k$ and the number of comparisons is small, LSD can be used with no problem.

As in any other linear model, it is necessary to do a residual analysis in order to check the hypothesis on the errors. The command `plot` works in the same way that for regression models (outputs from `lm` and `aov` are objects of class `lm`).

Effect of departures from the hypothesis

- Departures from normality
  If the departure is not severe, it has a limited effect (The F test is robust with respect to moderate departures from the normality hypothesis)
- Heterocedasticity
  If the groups have similar sizes, the effect of heterocedasticity is reduced.
- Independency
  The effect of the lack of independency can be enormous!